frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 25 April 2023
DOI 10.3389/fmars.2023.1174347

OPEN ACCESS

EDITED BY
Xuemin Cheng,
Tsinghua University, China

REVIEWED BY
Ning Wang,
Dalian Maritime University, China
Peng Ren,
China University of Petroleum (East China), China

*CORRESPONDENCE
Xiaodong Wang
wangxiaodong@ouc.edu.cn

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 26 February 2023
ACCEPTED 04 April 2023
PUBLISHED 25 April 2023

CITATION
Si G, Xiao Y, Wei B, Bullock LB, Wang Y and Wang X (2023) Token-Selective Vision Transformer for fine-grained image recognition of marine organisms. *Front. Mar. Sci.* 10:1174347. doi: 10.3389/fmars.2023.1174347

Token-Selective Vision Transformer for fine-grained image recognition of marine organisms

Guangzhe Si[1], Ying Xiao[2], Bin Wei[3], Leon Bevan Bullock[4], Yueyue Wang[5] and Xiaodong Wang[4]*

[1]College of Electronic Engineering, Ocean University of China, Qingdao, Shandong, China, [2]School of Science, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong SAR, China, [3]The Affiliated Hospital of Qingdao University/Shandong Key Laboratory of Digital Medicine and Computer Assisted Surgery, Qingdao University, Qingdao, Shandong, China, [4]College of Computer Science and Technology, Ocean University of China, Qingdao, Shandong, China, [5]Computing Center, Ocean University of China, Qingdao, Shandong, China

Introduction: The objective of fine-grained image classification on marine organisms is to distinguish the subtle variations in the organisms so as to accurately classify them into subcategories. The key to accurate classification is to locate the distinguishing feature regions, such as the fish's eye, fins, or tail, etc. Images of marine organisms are hard to work with as they are often taken from multiple angles and contain different scenes, additionally they usually have complex backgrounds and often contain human or other distractions, all of which makes it difficult to focus on the marine organism itself and identify its most distinctive features.

Related work: Most existing fine-grained image classification methods based on Convolutional Neural Networks (CNN) cannot accurately enough locate the distinguishing feature regions, and the identified regions also contain a large amount of background data. Vision Transformer (ViT) has strong global information capturing abilities and gives strong performances in traditional classification tasks. The core of ViT, is a Multi-Head Self-Attention mechanism (MSA) which first establishes a connection between different patch tokens in a pair of images, then combines all the information of the tokens for classification.

Methods: However, not all tokens are conducive to fine-grained classification, many of them contain extraneous data (noise). We hope to eliminate the influence of interfering tokens such as background data on the identification of marine organisms, and then gradually narrow down the local feature area to accurately determine the distinctive features. To this end, this paper put forwards a novel Transformer-based framework, namely Token-Selective Vision Transformer (TSVT), in which the Token-Selective Self-Attention (TSSA) is proposed to select the discriminating important tokens for attention computation which helps limits the attention to more precise local regions.

TSSA is applied to different layers, and the number of selected tokens in each layer decreases on the basis of the previous layer, this method gradually locates the distinguishing regions in a hierarchical manner.

Results: The effectiveness of TSVT is verified on three marine organism datasets and it is demonstrated that TSVT can achieve the state-of-the-art performance.

KEYWORDS

token-selective, self-attention, vision transformer, fine-grained image classification, marine organisms

1 Introduction

Fine-grained Image Classification (FIC) is a challenging task which utilizes subtle variations of the same species to differentiate the different subcategories, examples include birds (Van Horn et al., 2015), dogs (Khosla et al., 2011), and cars (Krause et al., 2013). Unlike general image classification, FIC requires sufficient attention being paid to the distinguishing features between the subcategories. There are a large number of highly similar fish and plankton in the ocean, and the classification of these subcategories (Li et al., 2019; Li et al., 2022) is conducive to the protection of marine ecology and biodiversity. However, the images of marine organisms are often taken in multi-angle and multi-scene situations, additionally, the background of marine life images is complex, which also increases the difficulty of recognition.

Recently, fine-grained image classification methods have made great progress due to the development of Deep Neural Networks (DNNs) (Simonyan and Zisserman, 2015; He et al., 2016; Liu et al., 2022; Shi et al., 2022; Wang et al., 2022). Strongly supervised fine-grained classification methods (Branson et al., 2014; Zhang et al., 2014; Wei et al., 2018) require labor-intensive labeling of images, so weakly supervised classification methods which rely only on category labels are now commonly preferred. CNN-based weakly supervised methods on fine-grained image classification can be mainly divided into localization methods and feature-encoding methods. Localization methods first locate the distinguishing regions and then extract features from these regions for classification. For example, some works (Ge et al., 2019; Liu et al., 2020) obtain the discriminating bounding boxes through Region Proposal Networks (RPNs) and then feed these regions into the backbone network for classification. However, the bounding boxes contain a lot of background areas with interfering information. Therefore, the discriminating regions localized by these methods are not precise enough. In addition, whilst the feature-encoding methods (Lin T.-Y. et al., 2015; Yu et al., 2018) make the output of the network change from semantic features to high-order features which can represent fine-grained information by means of feature fusion, the high-order features obtained by these methods have large dimensions, and the fine-grained information is not distinguishable.

Recently, Vision Transformer (ViT) (Dosovitskiy et al., 2021) has demonstrated potent performance on various visual tasks (Carion et al., 2020; Zheng et al., 2021; Guo et al., 2022). Specifically, in the task of image classification, a whole image is split into several patches, and each patch is converted into a token through linear projection. Then, the importance of each token is obtained through the Multi-Head Self Attention (MSA), and finally all of the tokens are combined according to the importance for classification. MSA in Transformer provides long-range dependency to enhance the interaction among image patches, so Transformer is able to locate subtle features and explore their relations from a large global scale perspective, whereas a traditional CNN has limited receptive fields and weak long range relationship abilities in very high layers with fixed-size convolutional kernels. ViT is therefore better suited to fine-grained classification tasks. In addition to the above advantages, ViT also has certain shortcomings, such as insufficient local sensing ability, tedious computation of MSA, and the need to consider the correlation among all tokens, our research is dedicated to improving these deficiencies.

Images of marine organisms are mostly taken from the bottom of the sea, the background of the images often contains reefs, corals and algae, which interferes with the recognition of the marine organisms themselves. A few images of marine life are taken from beaches, fishing boats and other scenes, the change of scenes also affects the identification of marine life. At the same time, due to the irresistible factors of camera angle and distance, images of the same subcategory show diverse global features, so paying too much attention to the global information is not conducive to correct classification. Examples of the three different scenarios are shown in Figure 1.

In this paper, to reduce the interference of intra-category diverse global information and useless background information, we propose a novel Token-Selective Vision Transformer (TSVT) for fine-grained image classification of marine organisms, which selects discriminative tokens layer by layer and gradually excludes interfering tokens. We propose a localized attention mechanism called Token-Selective Self-Attention (TSSA) to explore contextual information in discriminating regions and enhance the interaction amongst selected tokens. Influenced by the idea of clustering, for each discriminative token, only the other discriminative tokens related to it are selected for information interaction, then the class token integrates the information of these discriminative tokens for

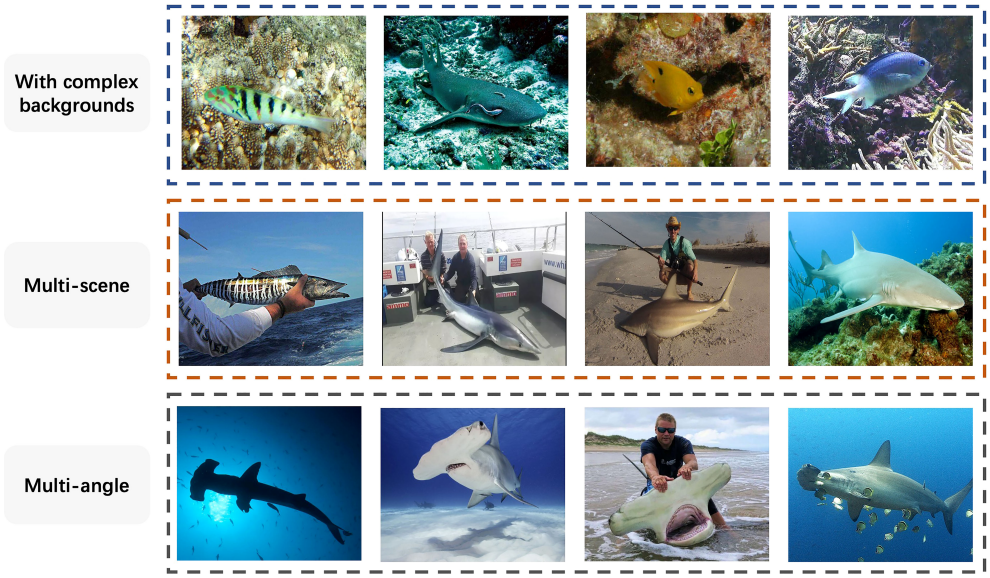

FIGURE 1
Some examples of marine life images. Three rows sequentially represent images with complex backgrounds, images of multiple scenes, and images of marine life taken from multiple angles.

classification. Finally, we verify the efficacy of TSVT for fine-grained image classification of marine organisms on three marine biological datasets.

In summary, our work has the following three contributions:

- We propose TSVT, a novel Vision Transformer framework for fine-grained image classification of marine organisms that excludes background interference and refines the range of distinguishing regions layer by layer.
- We propose Token-Selective Self-Attention (TSSA), which removes the interference of irrelevant tokens, and then establish the association of selected tokens in local regions and extract the most discriminative features.
- We conduct experiments on three different datasets to verify the effectiveness of our method, and show that TSVT achieves state-of-the-art performance. Additionally, we perform comparative experiments on TSSA's parameters to further explore the impact of applying TSSA to different layers, using different methods to select tokens and selecting different numbers of tokens on model performance.

2 Related work

2.1 Fine-grained image classification

2.1.1 CNN for fine-grained image classification

The fine-grained image classification methods based on CNN are mainly divided into two categories: localization methods and feature-encoding methods.

The basic idea of localization methods is to locate discriminative local regions first, and perform feature extraction on these regions, then cascade the extracted features and then again feed them to the sub-network for classification. Earlier localization methods (Zhang et al., 2014; Lin D. et al., 2015) rely on additional manual annotation information such as object bounding boxes and part annotation to help the network find the region with the most representative features. However, since such annotations are time-consuming and labor-intensive, more weakly supervised methods which only require image-level labels are preferred. Some methods (Ge et al., 2019; Liu et al., 2020) use RPN to obtain discriminative bounding boxes and input the selected feature regions into the network to capture local features. In addition, there are also methods to locate discriminative regions by utilizing an attention mechanism: RA-CNN (Fu et al., 2017) proposed Recurrent Attention to select a series of distinguishing regions for attention mapping in a coarse-to-fine manner; MA-CNN (Zheng et al., 2017) adopted a Multi-Attention CNN structure to obtain multiple distinguishing regions in parallel; MAMC (Sun et al., 2018) directed the generated attention features to categories to help better classification; NTS-Net (Yang et al., 2018) used a collaborative learning method to accurately identify the feature information regions.

Feature-encoding methods obtain richer fine-grained features for classification in the form of high-level feature interactions and the design of loss functions. As the most representative method for high-level feature interaction, B-CNN (Lin T.-Y. et al., 2015) used two deep convolutional networks to extract features from the same image, and then performed outer product operations on the feature vectors to obtain bilinear features for classification. However, the large feature dimensions of this method leads to a very large number of parameters, which is not easy to drive during training. To solve this problem, C-BCNN (Gao et al., 2016) adopted tensor sketches to reduce the dimensions of high-dimensional features.

Other methods attempt to capture features at higher levels to obtain a more distinguishable feature representation. HBP (Yu et al., 2018) combined the features of different layers through bilinear pooling, and finally concatenated them for classification. The loss function plays the role of a conductor's baton in Deep Learning and model learning is driven by it. In fine-grained image classification tasks, there are corresponding approaches to the design of loss functions: MaxEnt (Dubey et al., 2018) provided a training routine that maximizes the entropy of the output probability distribution; MC-Loss (Chang et al., 2020) focused on different local areas of each channel in the feature map, which is more conducive to feature learning.

2.1.2 ViT for fine-grained image classification

Transformer (Vaswani et al., 2017) was first applied to solve the sequence to sequence problem in Natural Language Processing (NLP) and has achieved better results than both convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Subsequently, Transformer has been widely used in the field of computer vision. ViT (Dosovitskiy et al., 2021) was the first transformer-based model for image classification, which splits images into a number of patches and inputs them to the transformer layer, and then establishes the association between different patches with the help of MSA, the classification is finally carried out by using the class token. TransFG (He et al., 2022) was the first to verify the effectiveness of vision Transformer on fine-grained visual classification. The input of its last layer is the class token and some important tokens representing distinguishing features rather than all of the tokens. In addition, RAMS-Trans (Hu et al., 2021) locates and extracts discriminative areas based on attention weights, and then re-inputs them into ViT for classification using multi-scale features.

In this paper, we propose TSSA, which allows each token to select its own relevant tokens according to the attention weights for attention computation. We integrate the one-to-one selection of each token into the attention computation. Furthermore, we apply TSSA to different layers of ViT to narrow the selection range layer by layer, so as to gradually refine the distinguishing features, yielding the major difference between our work and previous methods.

2.2 Underwater image classification

Due to the influence of the complex imaging environment in the ocean, the underwater images appear blurred, low contrast and low resolution, therefore various image preprocessing methods (Qi et al., 2022; Zhou et al., 2022; Zhou et al., 2023a; Zhou et al., 2023b) such as image enhancement and image restoration are used first to improve classification results. Recently, significant progress has been made in underwater classification, thanks to the influence of deep learning and the creation of several methods for underwater organism detection (Chen et al., 2021; Wang et al., 2023a; Wang et al., 2023b). The research on underwater biological image classification can be mainly divided into two aspects, one is the learning of biological features, the other is the feature fusion of different levels or types. For the feature acquisition methods, the earlier artificial methods (Alsmadi et al., 2010; Alsmadi et al., 2011) were only effective for specific datasets or scenarios, subsequently universal methods based on deep learning were adopted to learn various features. DeepFish (Qin et al., 2016) first extracted the fish regions using matrix decomposition, and then refined and learned these regional features by Principal Components Analysis (PCA) (Jackson, 1993) and CNN respectively. MCNN (Prasenan and Suriyakala, 2023) segmented fish images by the firefly algorithm and extracted features from the segmented parts. However, these methods require a large amount of computation, therefore, to maintain the balance between classification effect and cost, a number of efficient improved CNN networks were proposed: FDCNet (Lu et al., 2018) used filtering deep convolutional neural networks to classify deep-sea species; deconvolutional neural network was applied to different squid classification (Hu et al., 2020). In addition, in order to solve the noise background problem, AdaFish (Zhang et al., 2022) adopted adversarial learning to reduce the interference of background on classification.

Some methods (Kartika and Herumurti, 2016; Gomez Chavez et al., 2019) have obtained some limited improvement in classification accuracy by learning only a single feature such as fish color or coral texture, therefore combining multi-level or multi-part information to complete classification is another direction of underwater image classification. One method (Cui et al., 2018) integrated the texture and shape features of plankton to improve CNN performance; another method (Mathur et al., 2020) combined the characteristics of different parts of fish through cross convolutional layer pooling for prediction; whilst yet another method used a multi-level residual network (Prasetyo et al., 2022) which fused high and low level information through depth separable convolution was also proposed and achieved a good classification effect.

3 Methodology

3.1 Preliminary: vision transformer

The inputs of ViT are a sequence of serialized tokens. First, an image with resolution $H \times W$ is first split into fixed-size patches x_p, each of size $P \times P$, so the number of patches N is equal to $\frac{H}{P} \times \frac{H}{W}$. Each patch is transformed into a token x_{pt} by a patch embedding layer consisting of linear projection. In addition to patch tokens, there is a dedicated class token x_{cls} for final classification in the classification task. So all tokens include patch tokens and the class token. The above tokens only contain pixel information, and position encoding adds corresponding position information x_{pos} to each token to determine the position of each patch in the original image. All tokens are then fed into the transformer encoder, and the inputs of the transformer encoder x_0 are represented in Eq. 1:

$$x_0 = [x_{cls}; x_{pt}^1; x_{pt}^2; \ldots; x_{pt}^N] + x_{pos} . \quad (1)$$

Transformer encoder is the core of ViT and contains l transformer layers of MSA and Multi-Layer Perceptron (MLP) blocks, as well as residual connections after every block. The output of the l_{th} layer is represented as follows:

$$x_l^* = \mathrm{MSA}(\mathrm{LN}(x_{l-1})) + x_{l-1} \tag{2}$$

$$x_l = \mathrm{MLP}(\mathrm{LN}(x_l^*)) + x_l^*, \tag{3}$$

where $x_{l-1}$and x_l denote the encoded image representation of the $l-1_{th}$ and l_{th} transformer layers, x_l^* is the output of the MSA block after residual connection, LN represents layer normalization, and the class token of the last transformer layer is used for category prediction through MLP.

3.2 Overall architecture

Marine life images of the same subcategories present different global information such as posture and viewpoint, so an over-reliance on global information and a lack of attention to local information are not conducive to the correct classification. In addition, due to the complexity of the seabed environment, images of marine organisms often contain complex backgrounds such as reefs and corals, which will also affect the identification of marine organisms. In order to address the above issues, we first consider eliminating the interference of irrelevant factors such as the background, and locating the marine organisms themselves, then further locating the distinguishing areas. In this manner we propose TSVT, which selects tokens layer by layer for more accurate classification. By doing so, the number of tokens selected by the latter layer is further reduced on the basis of the preceding layer so as to more accurately refine the distinguishing areas and reduce the computational cost. To this end, we design a local attention module named TSSA, in which distinguishing tokens only interact with the other distinguishing tokens selected according to the attention weights, and the interference of background tokens is eliminated to obtain the purest distinguishing feature information for classification with the class token.

The framework of our TSVT is shown in Figure 2, where, the first eight transformers remain unchanged according to the settings of ViT, while the last four layers are Token-Selective Transformer Layer (TS Transformer Layer). It is different from the standard transformer layer in that it replaces the original MSA with TSSA. The number of tokens selected in each layer is different, and the local scope of attention is also different. The class token of the last layer aggregates the most discriminating features in the local regions and completes category prediction through MLP.

3.3 Token-selective self-attention

Fine-grained image classification requires focusing on local discriminating regions, but the complex background of marine biological images interferes with accurate localization of these regions. To solve the above issue, we propose to eliminate the interference of background tokens to the greatest extent and apply local attention to the selected important discriminating tokens.

All tokens can be divided into two categories: discriminating region tokens that play a positive effect in classification and background interfering tokens that play a negative effect in classification. Discriminating region tokens and background

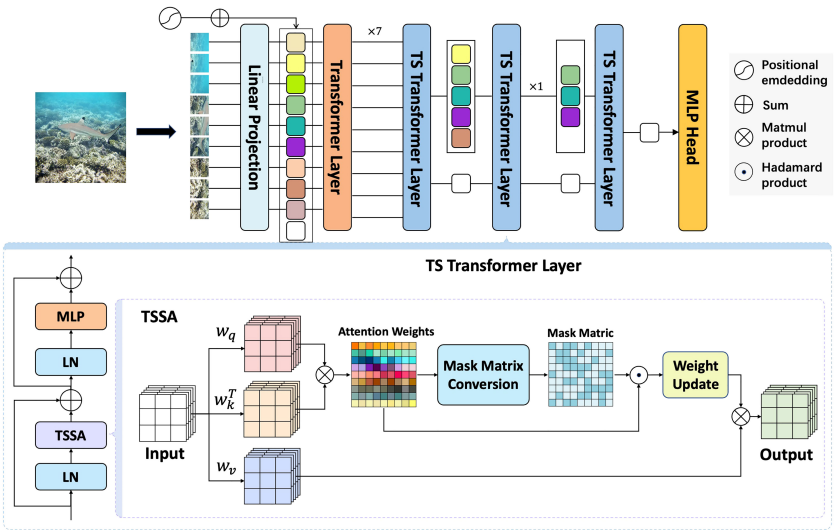

FIGURE 2
The framework of our proposed TSVT and the details of our designed TSSA. An image is first split into a number of patches, each of which is mapped into a feature vector by Linear Projection and combined with learnable position embedding. Contextual links between tokens are then established in the Transformer Layers, and the selection of tokens representing the discriminating regions is performed layer by layer in the latter four TS Transformer Layers with the number of selected tokens in each layer decreasing from the previous layers. In the TS Transformer Layer, TSSA is a sparse selective attention mechanism that generates a mask based on the similarity between tokens so as to limit the attention computation between non-relevant tokens.

tokens are clustered separately for information interaction in TSSA to ensure that discriminating tokens are no longer mixed with the interference information of background tokens, and then the class token integrates the information of distinctive tokens for the final classification.

The correlation between tokens can be reflected by attention weights. Previous work (Wang et al., 2021; He et al., 2022) has proved that attention weights can be a good indicator for token selection. The attention weights of each head in each transformer layer $A \in \mathbb{R}^{(N+1)\times(N+1)}$ can be written as follows:

$$A = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) = [a_0, a_1, a_2 \dots a_N], \tag{4}$$

$$a_i = [a_{i,0}, a_{i,1}, a_{i,2} \dots \dots a_{i,N}], i \in (0, N)\,. \tag{5}$$

According to the attention weights, the information of the token is weighted and summed to obtain the calculation result of the attention symbolized as *Attention*. The following formula is the calculation process of MSA:

$$Attention = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V, \tag{6}$$

where Q, K and V are all obtained by the linear transformations of tokens, all of which represent information about the token itself; d_k represents the dimensionality of K; softmax is a normalized exponential function; a_{ij} represents the degree of correlation between the i_{th} token and the j_{th} token, that is, token i as Q and token j as K for the calculation in Eq. 4; a_i represents the set of correlation degrees between the i_{th} token and all tokens; and $\cdot$ represents the general matrix product.

Only the largest m elements in each row of attention weights are selected, the selected elements remain unchanged, and the remaining unselected elements are all set to zero, thus generating new selective attention weights, which represent the degree of correlation between each token and its most relevant m tokens. In the computation of attention, the distinguishing tokens interact with each other and the distinguishing features are strengthened.

In the implementation, to ensure parallel computing, a mask matrix M with the same shape as the attention weights is first generated, we set the m_{th} largest element α_i in each row of attention weights as the threshold to determine whether the elements at different positions of mask matrix are one or zero. The process of mask matrix conversion is represented as:

$$M_{(i,j)} = \begin{cases} 1 & A_{(i,j)} \geq \alpha_i, \\ 0 & otherwise, \end{cases} \tag{7}$$

where (i, j) represents the position of each element in the mask matrix the and attention weights in $(n + 1) \times (n + 1)$ positions.

Then the selective attention weights A_s are obtained by computing the Hadamard product of the mask matrix and the attention weights, as follows:

$$A_s = A \odot M, \tag{8}$$

where $\odot$ is the calculation symbol for Hadamard product.

Without changing the relevance of the different tokens, we further update the elements a^s in the selective attention weights so that the sum of the elements in each row is equal to one, which further increases the proportion of discriminative information in the class token. Take the first row of A_s as an example, each element of this row $a^{s'}$ is computed as:

$$a^{s'}_{0,i} = \frac{a^s_{0,i}}{\sum_{j=1}^{N} a^s_{0,j}}\,. \tag{9}$$

The new selective attention weights $A_s^{'}$ represent the correlation between tokens in local areas, and then after the calculation in Eq. 10, the information between these tokens interacts and the output Z of TSSA is obtained. In the final TS Transformer Layer, the class token combines the token information through MLP for category predictions.

$$Z = A_s^{'} \cdot V\,. \tag{10}$$

The selective attention weights of each token-selective transformer layer are updated on the basis of the previous layer, and the number of selected tokens m of each layer is gradually reduced narrowing and refining the distinguishing feature regions layer by layer.

We apply TSSA to the deep layers of the model without destroying the globality of the shallow layers, and the local information based on the global basis is extracted for classification. Starting from the first token-selective transformer layer, the distinguishing tokens only aggregate important tokens related to them, so that the class token associated with these distinguishing tokens can minimize the interference of the background tokens. Our model is actually a trade-off between globality and locality, on the basis of not losing the globality, it can accurately locate the discriminating area and extract local features.

4 Experiments

In this section, we mainly introduce the experimental process and analyze the experiment results. First, we introduce the three marine biological datasets used in experiments, and briefly introduce the specific settings. Then, we verify the efficacy of TSVT by ablation study and analyze the experiment results.

4.1 Datasets

We validated the effectiveness of TSVT on three datasets of marine organisms, namely ASLO-Plankton (Sosik and Olson, 2007), Sharks[1], and WildFish (Zhuang et al., 2018). ASLO-Plankton consists of 22 categories of marine plankton images, its training set is unbalanced, and the number of images in different subcategories conforms to the long-tail distribution; Sharks contains images of 14 shark species, where the background of the images is complex and the differences between images are subtle; WildFish is a large-scale marine fish dataset with 1000 categories

and 54459 images in total, and we randomly select images of 200 categories from WildFish to form a new dataset WildFish200. The statistics of the three datasets are shown in Table 1.

4.2 Implementation details

The input image size of the ASLO-Plankton, WildFish200 and Sharks datasets is 448×448 pixels, the size of each patch is 16×16. We set the batch size on the three datasets to 8. SGD optimizer is employed with a momentum of 0.9. The learning rate is initialized as 0.03 and we adopt cosine annealing as the scheduler of optimizer. TSVT imports the pre-trained ViT-B_16 on ImageNet21k as the pretrained model. We complete the construction of the whole model using PyTorch and run all experiments on four NVIDIAGTX 1070 GPUs in one computer.

4.3 Comparison with the state-of-the-arts

Our method performs on par with a number of CNN-based methods: B-CNN (Lin T.-Y. et al., 2015), NTS-Net (Yang et al., 2018), TASN (Zheng et al., 2019), MC Loss (Chang et al., 2020), and the recent transformer variants: ViT (Vaswani et al., 2017), RAMS-Trans (Hu et al., 2021), TransFG (He et al., 2022) on ASLO-Plankton, Sharks and WildFish200. The experiment results are shown in Table 2. It can be seen from the results that ViT-based methods have a higher classification accuracy than CNN-based methods. Meanwhile, TSVT reaches 74.3%, 90.4% and 94.7% top-1 accuracy on ASLO-Plankton, Sharks and WildFish200 respectively, which achieves higher accuracy in the identification of marine organisms compared with other methods. The main reason for the improvement is that our method further eliminates background interference, accurately locates the discriminating areas, thus enlarging the differences between categories.

TABLE 1 Statistics of ASLO-Plankton, Sharks and WildFish200 datasets.

Dataset	Classes	Training	Testing
ASLO-Plankton	22	743	3300
Sharks	14	743	749
WildFish200	200	7929	3523

4.4 Ablation study

We verify the efficacy of our proposed TSSA on the three datasets, and further explore the impact of applying TSSA to different layers, using different methods to select tokens and selecting different numbers of tokens on model performance.

4.4.1 Impact of applying TSSA to different layers

We applied TSSA to the shallow layers (1-4), middle layers (5-8) and deep layers (9-12) of TSVT respectively, to explore the influence of token selection in different layers on model performance. The experiment results in the Table 3 show that applying TSSA to the deep layers achieves the best performance, whilst starting token selection in the shallow layers achieves worse performance. A possible reason is that the attention weights in shallow layers cannot highlight the key points that should be paid attention to, which is not enough to be used as the indicator for selecting tokens. On the contrary, with the deepening of layers, the feature information is accumulated, and the model starts to notice discriminating regions. At this time, further eliminating background and other interference can make the discriminative local features account for a larger proportion of final features used for classification. Global information needs to be strengthened by layers of accumulation, premature destruction of the association among all tokens at shallow layers is not conducive to extracting global features of the model. Therefore, establishing the association among all tokens at the shallow layers first, and then discarding some tokens at the deep layers is a trade-off between global information and local information, which is beneficial for classification.

When TSSA is applied to the deep layers, the classification performance of the model is improved. So we further explore the impact of applying TSSA to different deep layers. In different

TABLE 2 Comparison of TSVT and state-of-the-art methods on three datasets of marine organisms.

Method	Backbone	Accuracy(%)		
		ASLO-Plankton	Sharks	WildFish200
B-CNN	VGG-16	61.9	76.2	82.1
NTS-Net	ResNet-50	69.4	84.5	87.3
TASN	ResNet-50	70.0	85.2	88.7
MC Loss	ResNet-50	69.6	86.3	86.2
ViT	ViT-B_16	72.6	88.9	93.5
RAMS-Trans	ViT-B_16	73.1	89.2	93.8
TransFG	ViT-B_16	73.7	89.1	94.1
TSVT (Ours)	ViT-B_16	74.3	90.4	94.7

TABLE 3 Ablative experiments on applying TSSA to different layers.

Layers	ASLO-Plankton	Sharks	WildFish200
1-4	69.5	85.7	92.5
5-8	71.0	88.9	93.4
9-12	74.3	90.4	94.7

ablative experiments, the number of selected tokens decreases from the first TS transformer layer and the number in the final layer remains the same. As shown from the Table 4, the classification accuracy is constantly improved with the increase of the number of layers. The best effect is achieved when TSSA is applied to layers 8-12, which indicates that the model has been able to accurately locate the distinguishing regions from the 8_{th} layer, and the smaller the reduction of tokens between layers, the better the classification performance of the model.

4.4.2 Impact of the number of selected tokens

TSVT performs token selection layer by layer, and the latter layer continues to select tokens based on those selected in the previous layer in order to pinpoint discriminative regions hierarchically. In the experiments, we set a parameter p about the selection proportion to indicate the number of selected tokens, which is the ratio of the number of selected tokens to the number of all tokens. We studied the influence of the parameter p on the model, and the experiment results are shown in Table 5. When p is 0.7, TSVT achieves the best performance on the three datasets. As the p value increases from 0.7 to 0.9, the accuracy decreases, probably because too many background tokens are not discarded, leading to discriminative information being mixed with interference information. When the value of p is smaller than 0.7, the accuracy also decreases, which is because the number of tokens is too small and too many important tokens are discarded. When the value of p is smaller than 0.2, the number of selected tokens in the last layer is less than 1, so we did not conduct related experiments. In conclusion, TSVT is sensitive to the number of selected tokens.

4.4.3 Impact of token-selective methods

We select important tokens according to the attention weights. In this part, we select tokens randomly at layers 9-12 with the selection ratio $p = 0.7$ for comparison, which further verifies the efficacy of our selection method. The two methods of random selection and selection according to attention weights are respectively applied in TSSA for experiments. As can be seen from Table 6, the accuracy of the former method decreases by

TABLE 4 Ablative experiments on applying TSSA to different deep layers.

Layers	ASLO-Plankton	Sharks	WildFish200
12	73.2	88.9	93.7
11-12	73.6	89.4	94.3
10-12	73.4	90.0	94.3
9-12	74.3	90.4	94.7

TABLE 5 Ablation experiments on the number of selected tokens.

p	ASLO-Plankton	Sharks	WildFish200
0.9	73.2	89.1	93.8
0.8	72.9	89.4	94.4
0.7	74.3	90.3	94.7
0.6	72.9	89.9	94.3
0.5	72.1	88.5	93.7
0.4	71.3	88.1	91.1
0.3	69.2	87.9	88.7

3.6%, 1.6%, 0.6% respectively compared with the latter method (ours) on the three datasets. The reason is that some important distinguishing tokens are discarded in the process of random selection, and some tokens that interfere with classification accuracy may be selected for classification.

4.4.4 Visualization

In order to further verify the effectiveness of our method in locating discriminating regions, we use Grad-CAM (Selvaraju et al., 2017) to visualize the attention map generated from the attention weights of the final layer in TSVT and compare them with ViT. As shown in Figure 3, for images with complex backgrounds, ViT is easily affected by these backgrounds and focuses on objects irrelevant to classification, such as reefs and corals, while after excluding these interferences, TSVT easily locates marine organisms and their most distinctive features, such as patterns and spots on the fish. Taking the image in the first row and column as an example, ViT considers the human head as the discriminative region, while our method can accurately use the effective information of the hammerhead shark's head information to predict the category. In addition, for images where the fish are visually small due to the long shooting distance, TSVT can locate the positions of the small targets more accurately, whereas ViT sometimes cannot achieve such high precision positioning.

5 Conclusion

In this paper, in order to exclude the influence of the complex background of the seabed and accurately locate discriminating features, we propose a novel framework called TSVT for fine-grained image classification of marine organisms, which achieves the best performance on the three marine organism datasets compared with other state-of-the-art works. We propose a local attention mechanism called TSSA that excludes interfering tokens.

TABLE 6 . Ablative experiments on token-selective methods.

Selection Methods	ASLO-Plankton	Shark	WildFish200
random	70.7	88.8	94.1
max	74.3	90.4	94.7

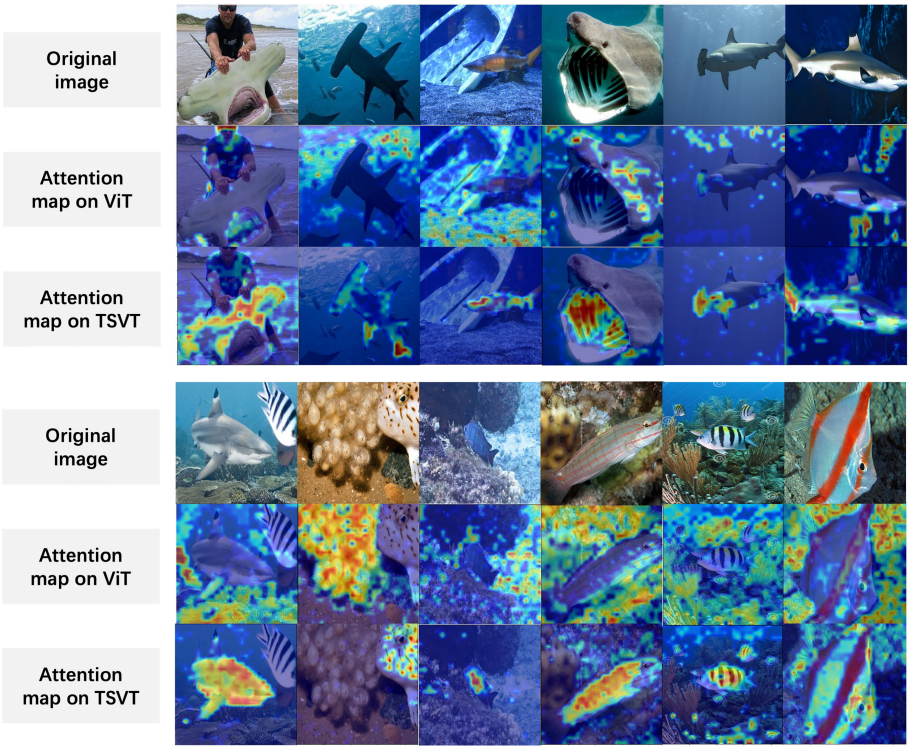

FIGURE 3
Visualization results on marine biological datasets, in which the first and fourth rows are six images in Sharks and WildFish datasets, the second and fifth rows are visualization of six images in the two datasets on ViT, and the third and sixth rows are visualization on TSVT.

Each discriminating token interacts with other discriminating tokens in the local area to extract positive fine-grained features to the greatest extent. Then, we explore the impact of applying TSSA to different layers, the number of selected tokens and token-selective methods on the performance of TSVT.

However, we still select key tokens through attention weights, which has the limitation that it must be applied to deep layers to ensure the reliability of the selection. Meanwhile, the number of key tokens in each image is not the same, so selecting tokens through more effective learning methods as well as setting learnable parameters to control the number of selected tokens is the future direction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

GS, YW, and XW designed the study and wrote the draft of the manuscript with contributions from YX and BW. BW and LB collected the marine fish image datasets. YW and XW devised the method. GS and YX performed the experiments. All authors contributed to the experimental analysis and manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32073029) and the Key Project of Shandong Provincial Natural Science Foundation (No. ZR2020KC027).

Acknowledgments

We thank the Intelligent Information Sensing and Processing Lab at Ocean University of China for their computing servers and collaboration during experiments. We kindly thank the Editor Dr. Xuemin Cheng for her efforts to handle this manuscript and all the reviewers for their constructive suggestions that helped us to improve our present manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alsmadi, M. K., Omar, K. B., Noah, S. A., Almarashdeh, I., Al-Omari, S., Sumari, P., et al. (2010). Fish recognition based on robust features extraction from size and shape measurements using neural network. *Comput. Sci.* 4, 1085–1091. doi: 10.3844/jcssp.2010.1088.1094

Alsmadi, M. K., Omar, K. B., Noah, S. A., et al. (2011). Fish classification based on robust features extraction from color signature using back-propagation classifier. *Comput. Sci.* 4, 52–58. doi: 10.3844/jcssp.2011.52.58

Branson, S., Van Horn, G., Belongie, S., and Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets. in. *Br. Mach. Vision Conference.* 2, 1–14.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. in. *Eur. Conf. Comput. Vision.* 2, 213–229. doi: 10.1007/978-3-030-58452-8_13

Chang, D., Ding, Y., Xie, J., Bhunia, A. K., Li, X., Ma, Z., et al. (2020). The devil is in the channels: mutual-channel loss for fine-grained image classification. *IEEE Trans. Image Process.* 4 (8), 4683–4695. doi: 10.1109/TIP.2020.2973812

Chen, T., Wang, N., Wang, R., Zhao, H., and Zhang, G. (2021). One-stage CNN detector-based benthonic organisms detection with limited training dataset. *Neural Networks* 4, 247–259. doi: 10.1016/j.neunet.2021.08.014

Cui, J., Wei, B., Wang, C., Yu, Z., Zheng, H., Zheng, B., et al. (2018). Texture and shape information fusion of convolutional neural network for plankton image classification. in. *OCEANS.* 5, 1–5. doi: 10.1109/OCEANSKOBE.2018.8559156

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: transformers for image recognition at scale," in *In International Conference on Learning Representations*, Vol. 2, 4. 1–22.

Dubey, A., Gupta, O., Raskar, R., and Naik, N. (2018). Maximum-entropy fine grained classification. in. *Adv. Neural Inf. Process. Systems.* 4, 1–11.

Fu, J., Zheng, H., and Mei, T. (2017). "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. in," in *IEEE Conference on Computer Vision and Pattern Recognition*, 3, 4438–4446.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016). "Compact bilinear pooling. in," in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 4. 317–326.

Ge, W., Lin, X., and Yu, Y. (2019). "Weakly supervised complementary parts models for fine-grained image classification from the bottom up. in," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2 (3), 3034–3043.

Gomez Chavez, A., Ranieri, A., Chiarella, D., Zereik, E., Babić, A., and Birk, A. (2019). CADDY underwater stereo-vision dataset for human–robot interaction (HRI) in the context of diver activities. *Mar. Sci. Eng.* 5, 1–14. doi: 10.3390/jmse7010016

Guo, Z., Gu, Z., Zheng, B., Dong, J., and Zheng, H. (2022). "Transformer for image harmonization and beyond," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2,

He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., et al. (2022). "TransFG: a transformer architecture for fine-grained recognition," in *In AAAI Conference on Artificial Intelligence*, 4 (6–8), 852–860.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition. in," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 770–778.

Hu, Y., Jin, X., Zhang, Y., Hong, H., Zhang, J., He, Y., et al. (2021). "RAMS-trans: recurrent attention multi-scale transformer for fine-grained image recognition," in *In ACM International Conference on Multimedia*, 4 (8), 4239–4248.

Hu, J., Zhou, C., Zhao, D., Zhang, L., Yang, G., and Chen, W. (2020). A rapid, low-cost deep learning system to classify squid species and evaluate freshness based on digital images. *Fisheries Res.* 4, 1–10. doi: 10.1016/j.fishres.2019.105376

Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 4, 2204–2214. doi: 10.2307/1939574

Kartika, D. S. Y., and Herumurti, D. (2016). "Koi fish classification based on HSV color space," in *In International Conference on Information Communication Technology and Systems*, Vol. 5. 96–100.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. (2011). Novel dataset for fine-grained image categorization:stanford dogs. in. *CVPR Workshop Fine-Grained Visual Categorization.* 2, 1–2.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). "3D object representations for fine-grained categorization. in," in *IEEE International Conference on Computer Vision*, 2, 554–561.

Li, J., Xu, W., Deng, L., Xiao, Y., Han, Z., and Zheng, H. (2022). Deep learning for visual recognition and detection of aquatic animals: a review. *Rev. Aquaculture* 2, 1–24. doi: 10.1111/raq.12726

Li, J., Xu, C., Jiang, L., Xiao, Y., Deng, L., and Han, Z. (2019). Detection and analysis of behavior trajectory for sea cucumbers based on deep learning. *IEEE Access* 2, 18832–18840. doi: 10.1109/ACCESS.2019.2962823

Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2015). "Bilinear CNN models for fine-grained visual recognition," in *IEEE International Conference on Computer Vision*, Vol. 2 (4–8), 1449–1457.

Lin, D., Shen, X., Lu, C., and Jia, J. (2015). "Deep LAC: deep localization, alignment and classification for fine-grained recognition," in *In IEEE Conference on Computer Vision and Pattern Recognition*, 3, 1666–1674.

Liu, C., Xie, H., Zha, Z.-J., Ma, L., Yu, L., and Zhang, Y. (2020). "Filtration and distillation: enhancing region attention for fine-grained visual categorization. in," in *AAAI Conference on Artificial Intelligence*, 2 (3), 11555–11562.

Liu, P., Zhang, C., Qi, H., Wang, G., and Zheng, H. (2022). "Multi-attention DenseNet: a scattering medium imaging optimization framework for visual data pre-processing of autonomous driving systems," in *IEEE Transactions on Intelligent Transportation Systems*, 2, 25396–25407.

Lu, H., Li, Y., Uemura, T., Ge, Z., Xu, X., He, L., et al. (2018). FDCNet: filtering deep convolutional network for marine organism classification. *Multimedia Tools Appl.* 4, 21847–21860. doi: 10.1007/s11042-017-4585-1

Mathur, M., Vasudev, D., Sahoo, S., Jain, D., and Goel, N. (2020).). crosspooled fishnet: transfer learning based fish species classification model. *Multimedia Tools Appl.* 5, 31625–31643. doi: 10.1007/s11042-020-09371-x

Prasenan, P., and Suriyakala, C. (2023). Novel modified convolutional neural network and FFA algorithm for fish species classification. *Combinatorial Optimization* 4, 1–23. doi: 10.1007/s10878-022-00952-0

Prasetyo, E., Suciati, N., and Fatichah, C. (2022). Multi-level residual network vggnet for fish species classification. *King Saud Univ. - Comput. Inf. Sci.* 5, 5286–5295. doi: 10.1016/j.jksuci.2021.05.015

Qi, Q., Li, K., Zheng, H., Gao, X., Hou, G., and Sun, K. (2022). SGUIE-net: semantic attention guided underwater image enhancement with multi-scale perception. *IEEE Trans. Image Process.* 4, 6816–6830. doi: 10.1109/TIP.2022.3216208

Qin, H., Li, X., Liang, J., Peng, Y., and Zhang, C. (2016). DeepFish: accurate underwater live fish recognition with a deep architecture. *Neurocomputing* 4, 49–58. doi: 10.1016/j.neucom.2015.10.122

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks *via* gradient-based localization," in *In IEEE International Conference on Computer*, 10, 618–626.

Shi, Z., Guan, C., Li, Q., Liang, J., Cao, L., Zheng, H., et al. (2022). "Detecting marine organisms *via* joint attention-relation learning for marine video surveillance," in *IEEE Journal of Oceanic Engineering*, 2, 959–974.

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition. in," in *International Conference on Learning Representations*, 3, 1–14.

Sosik, H. M., and Olson, R. J. (2007). Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology Oceanography: Methods* 8, 204–216. doi: 10.4319/lom.2007.5.204

Sun, M., Yuan, Y., Zhou, F., and Ding, E. (2018). "Multi-attention multi-class constraint for fine-grained image recognition. in," in *European Conference on Computer Vision*, 2, 805–821.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., et al. (2015). "Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. in," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 595–604.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. in. *Adv. Neural Inf. Process. Systems.* 4 (8), 1–11.

Wang, N., Chen, T., Liu, S., Wang, R., Karimi, H. R., and Lin, Y. (2023b). Deep learning-based visual detection of marine organisms: a survey. *Neurocomputing* 1–32, 4. doi: 10.1016/j.neucom.2023.02.018

Wang, H., Sun, S., Bai, X., Wang, J., and Ren, P. (2023a). "A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes," in *IEEE Journal of Oceanic Engineering*, 4, 1–19.

Wang, N., Wang, Y., and Er, M. J. (2022). Review on deep learning techniques for marine object recognition: architectures and algorithms. *Control Eng. Pract.* 118, 1–18 2. doi: 10.1016/j.conengprac.2020.104458

Wang, J., Yu, X., and Gao, Y. (2021). Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint arXiv* :2107.02341 6.

Wei, X.-S., Xie, C.-W., Wu, J., and Shen, C. (2018). Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition* 704–714, 2. doi: 10.1016/j.patcog.2017.10.002

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. (2018). "Learning to navigate for fine-grained classification," in *European Conference on Computer Vision*, 4 (8), 420–435.

Yu, C., Zhao, X., Zheng, Q., Zhang, P., and You, X. (2018). "Hierarchical bilinear pooling for fine-grained visual recognition," in *European Conference on Computer Vision*, 2 (4), 574–589.

Zhang, N., Donahue, J., Girshick, R., and Darrell, T. (2014). "Part-based r-CNNs for fine-grained category detection," in *European Conference on Computer Vision*, 2 (3), 834–849.

Zhang, Z., Du, X., Jin, L., Wang, S., Wang, L., and Liu, X. (2022). Large-Scale underwater fish recognition *via* deep adversarial learning. *Knowledge Inf. Syst.* 4, 353–379. doi: 10.1007/s10115-021-01643-8

Zheng, H., Fu, J., Mei, T., and Luo, J. (2017). "Learning multi-attention convolutional neural network for fine-grained image recognition," in *IEEE International Conference on Computer Vision*, 3, 5209–5217.

Zheng, H., Fu, J., Zha, Z.-J., and Luo, J. (2019). "Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8, 5012–5021.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2, 6881–6890.

Zhou, J., Sun, J., Zhang, W., and Lin, Z. (2023a). Multi-view underwater image enhancement method *via* embedded fusion mechanism. *Eng. Appl. Artif. Intell.* 4, 1–12. doi: 10.1016/j.engappai.2023.105946

Zhou, J., Yang, T., Chu, W., and Zhang, W. (2022). Underwater image restoration *via* backscatter pixel prior and color compensation. *Eng. Appl. Artif. Intell.* 4, 1–16. doi: 10.1016/j.engappai.2022.104785

Zhou, J., Zhang, D., and Zhang, W. (2023b). Cross-view enhancement network for underwater images. *Eng. Appl. Artif. Intell.* 4, 1–11. doi: 10.1016/j.engappai.2023.105952

Zhuang, P., Wang, Y., and Qiao, Y. (2018). "WildFish: a large benchmark for fish recognition in the wild," in *ACM International Conference on Multimedia*, 8, 1301–1309. doi: 10.1016/j.engappai.2023.105952